Highlighting Capability

Goal

When performing a search on the GCMD search portal the user is presented with a list of collections with the entry-id and a portion of the summary for that dataset. Within the summary any matches to the provided search terms will be highlighted. Also the part of the summary displayed is what is considered most relevant based on the user's search terms.

CMR will need to provide a way of returning the highlighted and most relevant portion of the summary.

Traceability

CMR-1480 - JIRA project doesn't exist or you don't have permission to view it.

CMR-1803 - JIRA project doesn't exist or you don't have permission to view it.

Links to either a JIRA epic, JIRA ticket, Jama L4, Jama User Story or Jama Review to show traceability between this design and the approved requirements driving this design

Design

CMR will use the Elasticsearch highlighting functionality documented here:

https://www.elastic.co/guide/en/elasticsearch/reference/current/search-request-highlighting.html

We will need to add a new query parameter 'include_highlighting' in both the parameter and JSON query search APIs. When returning the results we will return the highlighted summary snippet instead of the full summary field with each collection result.

Based on GCMD's responses to the questions below we will need to support additional parameters:

- highlight_snippet_length (default to full field)
- highlight_class (leaning towards not adding)
- highlight_begin_tag (defaults to)
- highlight_end_tag (defaults to)
- highlight_num_fragments (defaults to 1)

Those would be the names on the parameter API, it might be nice to provide a way to specify the parameters with JSON query, something like

```
{"highlights": {"snippet_length": 150,
    "class": "cmr_highlight",
    "begin_tag": "<b>",
    "end_tag": "</b>",
       "num_fragments": 2}}
```

Example

```
curl
"http://localhost:3003/collections.json?keyword=V1&include_highlighting=tr
ue&pretty=true"

;; Response
{
 "feed" : {
 "updated" : "2015-07-15T15:36:28.188Z",
 "id" : "http://localhost:3003/collections.json?keyword=V1",
 "title" : "ECHO dataset metadata",
 "entry" : [ {
  "score" : 0.5,
  "online_access_flag" : false,
  "id" : "C1200000000-PROV1",
  "browse_flag" : false,
  "summary" : "This is <em>V1</em> of the ET1 collection.",
  "original_format" : "ECHO10",
  "data_center" : "PROV1",
  "dataset_id" : "ET1",
  "title" : "ET1",
  "short_name" : "S1",
  "updated" : "2012-01-19T18:00:00.000Z",
  "orbit_parameters" : { },
  "version_id" : "V1"
  } ]
 }
}
```

Code Changes

- Initial prototype
  - Support include_highlighting query parameter (will support both parameter and JSON query API)
  - Add new highlight_results_feature namespace defining pre-process-query-result-feature and post-process-query-result-feature
- Additional features
  - Support the 5 highlighting parameters

Tests

- Verify highlighting works for each supported results format
- Verify highlights work with wildcards
- Verify multiple keyword searches (using AND/OR/NOT) result in multiple highlights as appropriate
- Verifications that each of the 5 highlighting parameters are handled correctly

Questions / Assumptions

1. Does the client need to be able to specify the tags used to identify the highlighting (default is match)?
   a. **GCMD: I think we should either allow passing &pre= &suffix=, or maybe include a text in the response format? Assume that would work with the tag...**
2. Do we need to support highlighting for fields other than summary? I think the answer is yes, but can pushed off to later sprint.
   a. **Will handle in future sprint.**
3. Does the client need to be able to specify the length of the summary snippet to return?

    a. **GCMD: We are happy if it is just similar length to what GCMD provides now.**
4. Should we return the entire summary field?
    a. **GCMD: As well? No, I don't think so**.
5. Does the client want more than one fragment to be returned?
    a. **GCMD: We have 2 fragments returned, that might be nice to be configurable.**
6. Assume that only keyword searches should be highlighted. Fielded searches would not have their terms highlighted in the summary field. (e.g. a keyword search for 'NSIDC' would highlight NSIDC within the summary, but a provider search for 'NSIDC' would not result in highlights in the summary field).
    a. **GCMD: Currently we just do "keyword"... I think it might be noisy if other things are included like facets and other fields.**

Error rendering macro 'pageapproval' : null